Harnessing Mixed Features for Imbalance Data Oversampling: Application to Bank Customers Scoring

Abdoulaye SAKHO[1,2], Emmanuel MALHERBE[1], Carl-Erik GAUTHIER[3], Erwan SCORNET[2]

[1]Artefact Research Center, [2] LPSM-Sorbonne Université, [3] Société Générale

ECML PKDD 2025

Introduction

Scoring: Addressing class imbalance in binary classification presents a significant challenge across various machine learning applications such as medical diagnosis, customer churn prediction or anomaly detection.

Standard techniques to handle class imbalance are mostly intrinsically designed for continuous input variables.

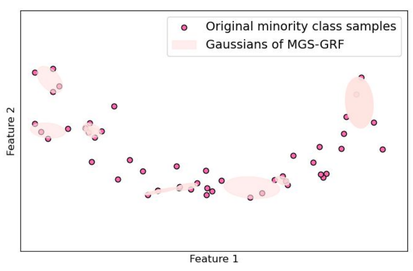
Private Data: millions of clients from Société Générale

Related Work

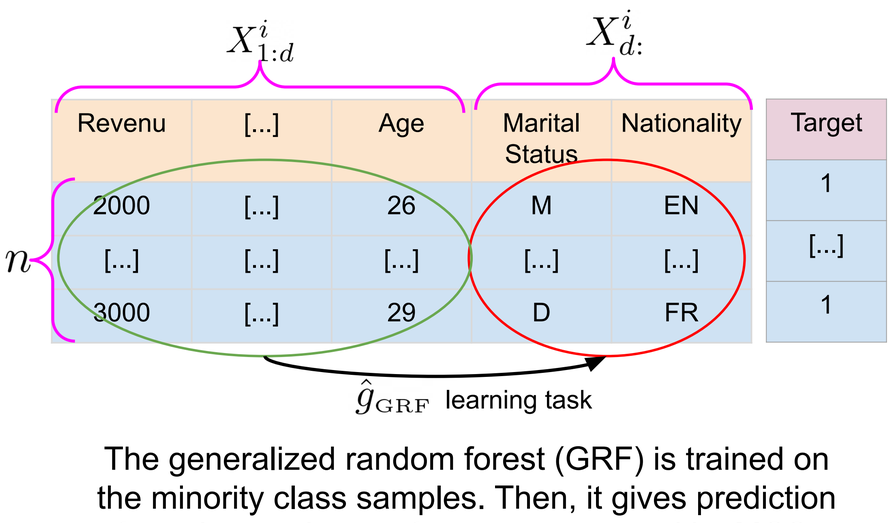
Oversampling strategies:

- Generate new synthetic samples within the minority class
- Generally *do not handle categorical features* by design
- Tends to copy sample and suffer from a boundary artifacts [1] (SMOTE-based)
- CTGAN [2] methodologies and ForestDiffusion [3] does not systematically outperform SMOTE with continuous input [see 1].

Proposed method: MGS-GRF



K = Number of nearest-neighbors =*d+1* for full rank covariance matrix



The generalized random forest (GRF) is trained on the minority class samples. Then, it gives prediction from the continuous features generated by MGS.

Generating continuous features:

$$\underbrace{\hat{f}_{MGS}(Z_{1:d})}_{\text{Kernel density estimator}} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi)^{d/2}|\hat{\Sigma}^i|} \exp\left(-\frac{1}{2}(Z_{1:d}-\hat{\mu}^i)^T(\hat{\Sigma}^i)^{-1}(Z_{1:d}-\hat{\mu}^i)\right)$$

$$\hat{\mu}^i = \frac{1}{K}\sum_{X\in \mathrm{NN}_{K,L_2}(X^i)} X_{1:d} \qquad \hat{\Sigma}^i = \frac{1}{K}\sum_{X\in \mathrm{NN}_{K,L_2}(X^i)} (X_{1:d}-\hat{\mu}^i)(X_{1:d}-\hat{\mu}^i)^T$$

Local kernel on samples neighbors

Generating categorical features:

$$Z_{d:} \leftarrow \hat{g}_{\mathrm{GRF}} \text{ samples } \{X^1_{d:},\dots,X^n_{d:}\} \text{ based on } \{w_{(Z_{1:d})}(X^1_{1:d}),\dots,w_{(Z_{1:d})}(X^n_{1:d})\}$$

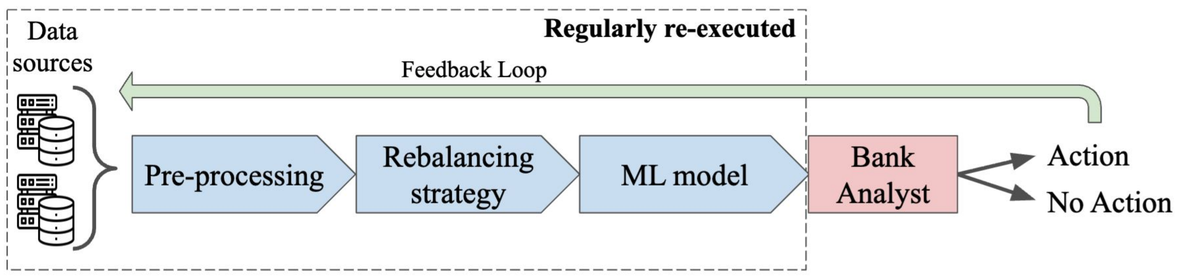
$$w_{(Z_{1:d})}(X^i) \leftarrow \frac{1}{T}\sum_{k=1}^{T} \frac{1_{\{X^i\in\mathcal{L}_k(Z_{1:d})\}}}{|\mathcal{L}_k(Z_{1:d})|}$$

GRF output weights for sample $Z_{1:d}$

Final algorithm to generate one sample:

$$\texttt{ContinuousSampler}(\{X^i_{1:d}\}_{i=1}^n) = Z_{1:d} \sim \hat{f}_{\mathrm{MGS}}$$
$$\texttt{CategoricalSampler}(\{X^i_{1:d}\}_{i=1}^n) = Z_{d:} \sim \hat{g}_{\mathrm{GRF}}(Z_{1:d})$$

Coherence of generated samples



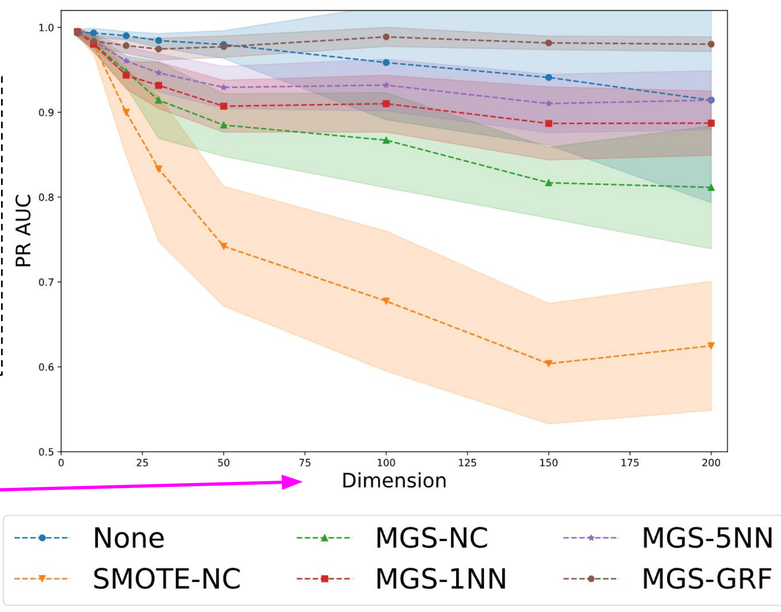
Business requirement: *Plausible* synthetic samples

$$Coh \triangleq \frac{1}{n_g}\sum_{\ell=1}^{n_g} I_{\{Z^\ell_{d:}\in\mathcal{C}_{Y=1}\}}$$

On synthetic data: the coherence *Coh* is positively correlated to better predictive performances.

Experiments on Simulated data

GRF preserves the dependance between continuous and categorical features, contrary to smote-based approaches



Results on real-world data sets with LightGBM

Metric	Data	Strategy							
		None	CW	ROS	RUS	SMOTE-NC	MGS -1NN	MGS -GRF	CW×MGS-GRF
PR AUC	Churn	0.622	0.608	0.576	0.394	0.595	**0.663**	**0.664**	-
	Mark.	0.092	0.090	0.090	0.082	0.076	0.098	**0.100**	-
	SocGen	*Ref.*	*+11%*	+7%	+10%	-28%	+10%	*+11%*	**+15%**
Time (s)	Churn.	0.30	0.33	0.49	0.06	0.85	1.22	0.89	-
	Mark	1.29	1.34	1.92	0.29	4.21	7.77	5.87	-

References

[1] Sakho A, Malherbe E, Scornet E. Do we need rebalancing strategies? A theoretical and empirical study around SMOTE and its variants. arXiv preprint arXiv:2402.03819. 2024 Feb 6.

[2] Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional gan. Advances in neural information processing systems. 2019.

[3] Jolicoeur-Martineau, A., Fatras, K. and Kachman, T., 2024, April. Generating and imputing tabular data via diffusion and flow-based gradient-boosted trees. In International conference on artificial intelligence and statistics (PMLR).